

SWITCHING DEVICE WITH MULTISTAGE QUEUING SCHEME

TECHNICAL FIELD

5 This invention relates generally to network switching devices. More particularly, this invention relates to a method and means for forwarding data packets through a switching device.

BACKGROUND OF THE INVENTION

10 A local area network (LAN) is a system for directly connecting multiple computers so that they can directly exchange information with each other. LANs are considered local because they are designed to connect computers over a small area, such as an office, a building, or a small campus. LANs are considered systems because they are made up of several components, such as cables,
15 repeaters, switches, routers, network interfaces, nodes (e.g., computers), and communication protocols. Ethernet is one such protocol. Information is communicated through a LAN in frames transported within data packets. ("Frame" and "data packet," while technically different, are often used interchangeably to describe data carrying the information.)

20 A LAN switch (or, more generally, a packet switch) is generally defined as a multi-port device that transfers data between its different ports based on the destination addresses and/or other information found in the individual packets it receives. Switches can be used to segment LANs, connect different LANs, or extend the collision diameter of LANs. Switches are of particular importance to
25 Ethernet-based LANs because of their ability to increase network diameter. Additional background information on packet switches can be found in a number of references such as *Fast Ethernet* (1997) by L. Quinn et al., *Computer Networks* (3rd Ed. 1996) by A. Tannenbaum, and *High-Speed Networking with LAN Switches* (1997) by G. Held, all of which are incorporated herein by reference.

30 Packet switches generally carry three types of traffic: unicast, multicast and broadcast. Unicast traffic consists of packets that travel from a source, or entry, port to a single destination, or exit, port. Multicast traffic consists of packets that

travel from one sending port to many destination ports per a destination list within a packet. Broadcast traffic is a special case of multicast traffic wherein the destination list includes all destination ports, and as such issues surrounding multicast apply equally to broadcast traffic.

5 Multicast traffic poses a problem for packet switches because multicast packets must be replicated within the packet switch. This replication can cause packet switches to fall behind in transmitting frames that follow the multicast frame due to the time required for replication. This replication time is particularly apparent in crossbar switch architectures that require unimpeded access from the
10 sending port to the destination ports to schedule transmission, as the sending port must wait for other traffic that is destined for the destination port to complete. Shared memory switch architectures do not suffer the same fate as all ports have access to the switch memory independent of each other. In shared memory switches, a packet is stored in a central memory and the sending port makes a
15 forwarding decision that notifies the destination ports of the packet's location for transmit. Each destination port can pull a multicast packet from its storage location independent of the other ports. However, in shared memory switches, the response of the destination ports to the forwarding decision can be time-consuming where, for example, it involves a multicast packet. Each of the destination ports on the
20 destination list must request the packet from the shared memory. The time required for this delays further forwarding decisions and can cause congestion in the sending port if additional traffic is received there while the destination ports complete their requests.

Congestion in packet switches can be caused in many ways. In cases where
25 there is more than one port transmitting to a single destination port, congestion at the destination port can occur and the port is said to be oversubscribed. The ratio of the rate that traffic is generated at the sending ports to the rate the destination port can transmit is called the oversubscription ratio. Rate mismatches in the source network media and destination network media can also cause congestion
30 (another case of oversubscription). For example, if traffic travels from port 0 to port 1 on a switch and port 0 runs at 100 megabits per second and port 1 runs at 10 megabits per second, traffic can easily back up waiting to exit port 1. Traffic

0916343 100598

shaping can also cause congestion. This is a process where the destination port is intentionally curbed back to a lower transmit rate than it is capable for traffic engineering purposes.

Normal packet network traffic does not allow steady state oversubscription.

- 5 No network switching equipment can buffer infinite data, so oversubscribed conditions on ports are inherently limited by the storage capabilities of the network equipment. However, it is also normal network behavior to have bursts for short periods of time during which network ports will be oversubscribed for any or all of the reasons previously listed. High port count switching equipment creates the
- 10 opportunity for high oversubscription ratios during these normal traffic bursts.

- End stations (such as a node within a LAN) communicate through a packet network by establishing a channel called a session. This session has characteristics that remain constant during the conversation between the end stations. For example, if station A talks to station B through the packet network to transfer a
- 15 file, when A sends packets, they are all labeled with B's network address as the destination, and A's network address as the source. Other information in the packets sent between A and B will also generally remain constant for a given session - priority, VLAN, network protocol, etc. Each station in a packet network may run multiple sessions with the same or different destination stations. In
- 20 general, packets within these sessions must arrive in sequence at their destinations - that is, the network equipment must not re-order them.

- As a packet from a given session enters a packet switch, the switch must evaluate fields within the packet and make a forwarding decision (where does the packet go?). After making a forwarding decision, packet switches with egress
- 25 queuing must place multicast packets on more than one transmit queue (a queue being first in/first out storage). As network media speeds increase, the time allowed to perform these queuing operations shrinks. Egress port congestion aggravates this issue. High oversubscription ratios that result from normal packet network operation force egress queuing mechanisms to queue packets from many
- 30 sources simultaneously to maintain predictable operation. The more work the queuing mechanism performs to handle congestion, the harder it is to perform multicast packet replication.

An objective of this invention, therefore, is to provide a method and means for enhancing the communication of packets, such a multicast packets, through a switching device.

5

SUMMARY OF THE INVENTION

In a switching device, a method of communicating data packets from sending ports to destination ports includes storing in a first stage queue packet-related data from a sending port; determining from the packet-related data which destination ports are to receive the packet-related data in the first stage queue; storing in a second stage queue associated with each determined destination port the packet-related data from the first stage queue; and using the packet-related data in the second stage queue to complete the communication of the data packet from the sending port to each determined destination port. Apparatus for practicing the method comprises a first stage queue storing packet-related data from a sending port; and a second stage queue associated with each of a set of destination ports storing the packet-related data from the first stage queue.

These and other aspects, features, and advantages of the invention are described in an illustrative embodiment below in conjunction with the following drawings.

20

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is an overall block diagram of a packet switch in accordance with the invention:

Fig. 2 is a block diagram of a queuing device within the packet switch of Fig. 1.

25

DETAILED DESCRIPTION OF AN ILLUSTRATIVE EMBODIMENT

Overview

As a session's packets travel through the packet network, they may take different paths to the same destination based on data field contents in their packets. For example, a video conference session between station A and B may be marked as priority 7, VLAN 5 while an email session between the same two stations may

30

be marked as priority 3, VLAN 12 even though the sessions have the same source and destination. As a result of these differences in the sessions, the packets in the data streams may take physically different routes through the network. The sessions may also take the same route through the network, but be serviced
5 differently by the network equipment - a packet from one session may come into a switch after a packet from a different session, yet exit the switch first as a result of higher priority servicing.

Network switches can examine these data fields in each packet to classify it. This invention takes advantage of packet classification to reduce the bandwidth
10 burden of queuing multicast packets during congested intervals. This is accomplished by identifying classifications (characteristics) of packets that are independent of whether a packet is unicast or multicast. By doing this, queuing can be broken up into stages such that the first stage is only concerned with a broad range of packet types (and potentially sessions) that are destined for any port in a
15 group of ports. Further stages can then service these queues as appropriate and perform additional queuing for multicast by replicating the packet for each destination port. The multicast replication still occurs within the switch, but is not required to occur immediately as the first stage of queuing acts as a buffer for the subsequent stages. This allows for the absorption of normal bursty
20 oversubscription conditions while maintaining session servicing consistency (packet ordering).

An example of this invention is a two stage queuing scheme whereby the first stage queues packets based solely on their priority. All sessions with a given priority are given equal weight in the first stage queue. Note that these include
25 multicast and unicast sessions. All packets within a session are placed in a single queue, thus maintaining packet ordering requirements. A second stage queue services or empties the priority queues into port queues, expanding multicast packets onto multiple queues if necessary. Since the first stage queues are not required to expand multicast packets onto multiple destination ports, fewer queuing
30 operations are required to queue the packets to their destinations in the first stage. Multicast replication is handled by second stage queuing mechanisms. This allows more ports to be effectively utilized in normal bursty oversubscribed packet

09166343-100598

network conditions due to rate mismatches, traffic shaping, oversubscription and port density. This is increasingly important as network media speeds achieve gigabit per second and beyond.

The operation of the illustrative embodiment is such that the multistage queuing is not visible except during congested intervals on ports receiving multicast traffic. That is, it is transparent during normal operation and operates as a performance enhancement during peak utilization.

Note that this invention is not limited to the use of priority as a first stage queuing criteria. This is an arbitrary session characteristic that was chosen as a classification rule, which could as easily have been network protocol type or type of service. Note also that this invention is not limited to egress queuing devices, and may have applications in ingress queuing packet switches. Nor is this invention limited to a two stage queuing pipeline.

15 *Structure and Function*

Fig. 1 is a block diagram of a portion of a switching device 18 in accordance with the invention. The switching device includes a number of bidirectional ports 20 (numbered individually as ports 1 through 10) each which includes a media access control (MAC) and forwarding decision logic 22. For purposes of this description, a port that acquires a data packet from an external entity such as a network, node, station, etc., for example, and forwards the packet internally to another port is referred to as an entry or sending port. A port that receives a data packet internally and transmits the data packet to an external entity is referred to as an exit or destination port. Also shown within the switch 18 is a switch fabric 24 and a queuing device 26. Other portions of the switching device, which can be conventional in nature, are not shown and are not described because they are not germane to the invention.

The MAC and forwarding decision logic 22 within each port can be conventional in design. Through them a sending port acquires a data packet and forwards it to one or more destination ports. The process for communicating forwarding decision is made by the forwarding decision logic based on destinations listed in the data packet. The forwarding decision logic 22 forwards to the queuing

0916343-100598

device information (packet-related data) such as a pointer to where the packet is stored in the switch fabric (if the fabric is shared memory), the type of packet (priority, etc.) and to which ports the packet is to be communicated (one port for unicast, multiple ports for multicast, and all but the sending port for broadcast).

- 5 The forwarding decision logic also stores the data packet in the switch fabric 24 at the location indicated by the pointer.

The switch fabric 24 in the illustrative embodiment is a shared memory that stores the entire data packet and from which destination ports may retrieve a copy. However, the invention is not limited to shared memory architectures. The switch
10 fabric may be of other architectures such as a crossbar matrix.

The queuing device 26, which receives the packet-related data from the forwarding decision logic 22, is shown in more detail in Fig. 2. The device 26 includes characteristic detection logic 30, a first stage queue 32 coupled to the output of the detection logic, port membership determination logic 34 coupled to
15 the output of the first stage queue, and second stage logic 36 coupled to the output of the determination logic.

The characteristic detection logic 30 detects from the packet-related data a characteristic of the packet (e.g., priority) as well as determining from a destination list which ports are to retrieve the packet. The destination list in the
20 illustrative embodiment is encoded in the packet-related data as a number, which is used by the logic 30 to look up the associated port group in a look up table. Other means may also be used for indicating the destination ports, such as passing the port numbers directly within the packet-related data.

The first stage queue 32 stores the packet-related data according to a
25 characteristic of the packet. In the illustrative embodiment, that characteristic is the priority of the packet (from one to eight levels). Consequently the first stage queue includes multiple first queues, each one storing the packet-related data (pointer plus port group) for packets of a different priority level in the order in which they are received.

30 The port determination logic 34 reads the first stage queue according to a service scheme and determines from the packet-related data which destination ports

09166343-100599
065007-5499T60

are to receive the packet-related data in the first stage. The logic 34 then stores the pointers in the appropriate portion of the second stage queue 36.

The second stage queue 36 in the illustrative embodiment includes for each port connected to the queuing device a set of queues that correspond to the first stage queue. For example, the first stage queue includes eight priority queues, and the second stage queue includes for each of its connected ports a corresponding set of eight priority queues. With this arrangement, the pointers in each queue of the first stage queue can be easily copied into a corresponding queue in the second state queue for each determined destination port.

The queuing device also includes conventional logic (not shown in Fig. 2 for clarity) for requesting packets from the switch fabric 24 (packet req.) and for forwarding packets from the switch fabric to the destination ports (xmt packet). This conventional logic is included in the queuing device in the illustrative embodiment as an implementation choice. However, the logic could just as well be separate from the queuing device if desired.

Operation

A data packet communicated from a sending port of switch 18 to one or more destination ports travels through the switching device as follows. A data packet that is received by a port is processed by its MAC, which generates a well formed frame from the physical LAN interface. The MAC presents the packet to the forwarding decision logic 22 that classifies the packet type and makes a forwarding decision as to which ports the packet is to be sent to from a destination list within the packet. The forwarding decision logic 22 transfers the packet to a location in the switch fabric (shared memory in the illustrative embodiment). The forwarding decision logic 22 also generates the packet-related data which includes the type of packet, a pointer to the location in shared memory where the packet is stored, and the forward decision and sends this data to the queuing device 26.

The queuing device receives the packet-related data at detection logic 30 and stores it in the first stage queue 32. In the process of storing the data in the first stage queue the logic determines which from the data ports, if any, the packet should be queued for. The logic then stores the pointer and destination port

09165343-100598

information in an appropriate queue within the first stage queue. In the illustrative embodiment the data is stored in a queue based on the priority of the packet; other characteristics of a packet can also be used for determining where the data will be stored.

5 The queuing device, through the determination logic 34, then obtains data from the multiple queues of the first stage in accordance with a scheme for reading the queues, such as a priority servicing scheme where the characteristic is priority. The determination logic 34 determines from the packet-related data just obtained which destination ports are to receive the pointer(s) and transfers the pointers to an
10 appropriate location in the second stage queue 36 associated with each determined destination port. In the illustrative embodiment the second stage queue includes multiple queues for each port that correspond to the multiple queues of the first stage queue. In this embodiment, with priority as the packet characteristic, the logic 36 transfers the pointers to the priority queue for each port corresponding to
15 the priority queue for the pointer in the first stage queue.

 The queuing device then uses the data in the second stage queue to complete the communication of the data packet from the sending port to each determined destination port. When servicing the second stage queue with a service scheme, the queuing device obtains pointers from the queues to packets in the
20 switch fabric 24 and requests these packets from the switch fabric. The switch fabric responds by sending the pointed-to packet to the queuing device, which then directs it to the appropriate port. Where there are multiple destination ports such in the case of a multicast data packet, the queuing device makes separate requests for each port.

25 The queuing device 26 is shown in the output or transmit path of the switch 18. In this configuration, the switch is said to be a "transmit buffered" or "output buffered" device. For different switch architectures like crossbar, the queuing device may reside on the inbound side of the switch, but the invention may still be applied. In that application, it would be referred to as "input buffered" since the
30 queues are stored at the sending ports of the switching device.

 Having understood the principles of the invention from the embodiments of the invention shown and described herein, those of skill in the art will recognize

09166343-100598
"SECRET"

that the embodiments can be modified in arrangement and detail without departing from such principles. The construction of the various modules can be varied while still providing the functions described. Elements of the various modules can be implemented in hardware, software, or firmware as desired. The packet-related

5 data may be pointers, other structures, or the data packets themselves. The invention may be used where appropriate in any packet switching device such as a LAN switch, a router, etc.

In view of the many possible embodiments to which the principles of the invention may be applied, it should be understood that these embodiments are

10 illustrative only and should not be taken as a limitation on the scope of the invention. The invention, rather, is defined by the following claims. We therefore claim as the invention all embodiments that may come within the scope of these claims and their equivalents.

09166343-100593
3500T "CHE9T60